2019-03-12

HUMAN BRAIN MAPPING

The effect of mental countermeasures on neuroimaging-based concealed information tests

Journal:	*Human Brain Mapping*
Manuscript ID	HBM-18-0798.R2
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Hsu, Chun-Wei; University of Plymouth, University of Plymouth, School of Psychology; University of Plymouth, Cognition Institute Begliomini, Chiara; University of Padova, Department of General Psychology; University of Padova, Cognitive Neuroscience Center Dall'Acqua, Tommaso; University of Padova, Department of General Psychology Ganis, Giorgio; University of Plymouth, School of Psychology; University of Plymouth, Cognition Institute
Keywords:	neuroimaging, concealed knowledge, memory detection, countermeasures

The effect of mental countermeasures on neuroimaging-based concealed information tests

Chun-Wei Hsu[1], Chiara Begliomini[2,3], Tommaso Dall'Acqua[2], & Giorgio Ganis[1]

1. School of Psychology and Cognition Institute, University of Plymouth, Portland Square, Drake Circus, Plymouth, PL4 8AA, UK
2. Department of General Psychology, University of Padova, Via Venezia 8, Padova, Italy
3. Cognitive Neuroscience Center, University of Padova, Padova, Italy

Corresponding author: Giorgio Ganis, giorgio.ganis@plymouth.ac.uk

Acknowledgments

We would like to acknowledge Prof Clara Casco and Dr. Luca Battaglini for invaluable help with the logistics of running this study.

This study was funded by a fellowship from the People Programme (Marie Curie Actions) of the European Union's CogNovo (*FP7-PEOPLE-2013-ITN-604764*).

Abstract

During the last decade and a half, functional magnetic resonance imaging (fMRI) has been used to determine whether it is possible to detect concealed knowledge by examining brain activation patterns, with mixed results. Concealed information tests rely on the logic that a familiar item (probe) elicits a stronger response than unfamiliar, but otherwise comparable items (irrelevants). Previous work has shown that physical countermeasures can artificially modulate neural responses in concealed information tests, decreasing the accuracy of these methods. However, the question remains as to whether purely mental countermeasures, which are much more difficult to detect than physical ones, can also be effective. An fMRI study was conducted to address this question by assessing the effect of attentional countermeasures on the accuracy of the classification between knowledge and no-knowledge cases using both univariate and multivariate analyses. Results replicate previous work and show reliable group activation differences between the probe and the irrelevants in fronto-parietal networks. Critically, classification accuracy was generally reduced by the mental countermeasures, but only significantly so with region of interest analyses (both univariate and multivariate). For whole-brain analyses, classification accuracy was relatively low, but it was not significantly reduced by the countermeasures. These results indicate that mental countermeasure need to be addressed before these paradigms can be used in applied settings and that methods to defeat countermeasures, or at least to detect their use, need to be developed.

Highlights

- FMRI-based concealed information tests are vulnerable to mental countermeasures
- Measures based on regions of interest are affected by mental countermeasures
- Whole-brain analyses may be more robust than region of interest ones
- Methods to detect mental countermeasure use are needed for forensic applications

Keywords

Neuroimaging, concealed knowledge, memory detection, countermeasures

Introduction

Concealed information tests (CITs) are used to determine if an individual is familiar with a certain piece of information such as a crime-related item (Ben-Shakhar, 2012; Ganis, Rosenfeld, Meixner, Kievit, & Schendan, 2011; Meijer, Selle, Elber, & Ben-Shakhar, 2014; Rosenfeld, Ben-Shakhar, & Ganis, 2012). The basic logic of these tests (Lykken, 1959) is that recognition of an item of interest (usually a crime-related item, referred to as "probe") will generate a stronger response compared to suitable control items (referred to as "irrelevants") that can be measured by monitoring behavioral, psychophysiological, or neural variables. To illustrate, the perpetrator of a crime should show a stronger response to a crime item (e.g., the crime weapon) than to other items that could have been used to commit the crime but weren't (e.g., other possible weapons) because the perpetrator would recognize the crime item but not the other items. In contrast, a person who was not involved in the crime should show comparable responses to all items. Thus, the difference in response between the probe and the irrelevants ("probe effect", hereafter) can be used as an index of whether somebody has concealed knowledge about a certain item.

Currently, the CIT (using psychophysiological variables) is systematically employed for forensic purposes only in Japan, but there has been growing interest in it across the world (Matsuda, Nittono, & Allen, 2012; Verschueure, Ben-Shakhar, & Meijer, 2011). Among the important issues for any potential application of the CIT is the extent to which countermeasures, physical or mental, can reduce its accuracy (Honts, Devitt, Winbush, & Kircher, 1996). An effective physical countermeasure with polygraphy-based CITs involves pressing one's toes to the floor during the presentation of irrelevants (Honts et al., 1996). Similarly, an effective mental countermeasure entails counting backwards by sevens to irrelevants (Honts et al., 1996). Countermeasures can be effective even with neuroscience-based methods that at first sight may seem more difficult to compromise. With event-related potentials (ERPs), a physical countermeasure that relies on associating covert actions with irrelevants to increase their relative saliency can reduce the size of the probe effect (Rosenfeld, Soskins, Bosh, & Ryan, 2004), prompting researchers to try to develop CIT variants that are less vulnerable to countermeasures (Bowman et al., 2013; Rosenfeld et

al., 2008). This countermeasure can also disrupt the accuracy of a common kind of fMRI-based CIT (Ganis et al., 2011), most likely by affecting the saliency of the CIT items, and by modulating activation in the salience network (Seeley et al., 2007) engaged by the CIT. Note that although this countermeasure is technically a physical one because it involves making specific "imperceptible" movements with one's fingers and toes in response to a subset of irrelevants, it is not a physical countermeasure in the classic sense. This is because it was not designed to produce the kind of changes in autonomic nervous system activation (e.g., the skin conductance or heart rate changes produced by painful self-stimulation) that make physical countermeasures effective (Honts et al., 1996). Rather, the countermeasure was designed to make some irrelevant items more salient by associating them with covert actions.

Although our study was the first on this topic (Ganis et al., 2011), it had some limitations. First, it examined only a very specific type of countermeasure, leaving open the issue of whether a more general class of mental countermeasures may also be effective (Ganis, 2018a). Second, the countermeasure was applied only to a subset of the irrelevants because applying it to all irrelevants would have artificially increased the saliency of the probe, being this the only item without an associated countermeasure. This may have not only diluted the effect of the countermeasure, as one typically compares the probe with the mean of all the irrelevants, but it may also provide clues of countermeasure use. Thus, countermeasures that can be applied to all items may be more effective. Third, this countermeasure was detectable by examining activation in primary motor cortex, as it involved making irrelevant-specific imperceptible movements with one's fingers and toes and so it engaged motor planning and motor imagery.

An interesting class of mental countermeasures that may address some of the limitations just discussed was tested in recent fMRI work in a different context using standard old/new face recognition paradigms (Rissman, Greely, & Wagner, 2010; Uncapher, Boyd-Meredith, Chow, Rissman, & Wagner, 2015). This work showed that multivariate analyses of brain activation can discriminate well above chance hits (correctly recognized old faces) from correct rejections (correctly rejected new faces) in single individuals. However, the accuracy of the discrimination was reduced to chance by using attentional and memory

countermeasures because they altered activation in regions involved in memory retrieval and goal-directed shifts of attention (Uncapher et al., 2015). On the one hand, patterns of brain activation associated with a new face could be made to resemble that of an old (recognized) face by bringing to mind a known individual that resembled that face, reliving memories associated with that individual, and by responding as if it was an old face (i.e., pressing the "old" key). On the other hand, brain activation associated with an old face could be made to look like that of a new face by diverting attention away from the recognition experience and by focusing instead on technical and photographic aspects of the old face (e.g., exposure, lightning, and so on) and by responding as if it was a new face (i.e., pressing the "new" key).

There are notable differences between standard recognition paradigms and CITs used in fMRI studies. Specifically, standard recognition paradigms usually employ hundreds of stimuli that are presented only once during study and test and a new stimulus in this paradigm is never encountered before in the study. In contrast, CITs used in fMRI studies typically employ fewer than 10 stimuli and these stimuli are repeated tens of times during testing: irrelevant stimuli in these paradigms are not new as the new items in recognition paradigms, in the sense that they have been encountered many times before in the CIT session, like the probes. Therefore, the neural processes involved in discriminating old and new items in standard recognition paradigms are likely to be different from those involved in discriminating between probes and irrelevants in CITs. Despite these differences, we predicted that attentional and memory countermeasures of this kind would also be effective with CITs. Thus, we devised a novel mental countermeasure with two components, one requiring attention to be focused on superficial features of the probe in order to decrease its saliency, and the other to reactivate meaningful memories associated with the irrelevants to increase their saliency. Although this countermeasure is similar to the one used by Uncapher and collaborators (Uncapher et al., 2015), it is not identical to it due to the different stimuli and paradigms used in the two studies. Hereafter, we will refer to this countermeasure in the singular or in the plural ('countermeasures', since there are two components), depending on the context.

In sum, we compared the probe effect in concealed knowledge, no knowledge, and countermeasure conditions and used both univariate and multivariate analyses to determine

the effect of mental countermeasures. We expected to find a reliable probe effect in the same prefrontal-parietal network reported in previous CIT studies in the concealed knowledge condition (Ganis et al., 2011; Peth et al., 2015), compared to the no knowledge condition, which in turn was expected to show no probe effect. Furthermore, in the countermeasures condition we expected the probe effect to be smaller and classification accuracy to be lower than in the concealed knowledge condition.

Materials and methods

Subjects

Twenty-three right-handed normal individuals (9 females; mean age = 24.2 years) were recruited among Padova University students participated in the study. Exclusion criteria included history or presence of neurological or psychiatric disorders and failure to meet one or more of the screening criteria regulating MRI scanning safety procedures. Three participants did not press any button on more than 50% of the trials and their data were not used. All analyses were carried out on the remaining twenty participants (9 females, mean age = 24.5 years). The Ethics Committee of Padova University Hospital approved the study and all participants gave signed informed consent before taking part in the experiment.

Stimuli

In each task, the stimuli were six digits (3-8) shown in white against a black background and presented for 750 ms. The stimuli were followed by a black screen with a fixation dot lasting between 1000 and 9000 ms (2000 ms on average), according to a pseudo-random sequence (Dale, 1999). These stimuli were used because they were very similar to each other visually and they had already been successfully used in prior work by this group (Ganis, Bridges, Hsu, & Schendan, 2016).

Design and Procedure

The study took place at the Radiology Unit of Padova University Hospital where

fMRI scanning was conducted.

Stimuli for the event-related fMRI tasks were presented using E-prime 2.0 software (Psychology Software Tools, Inc., Sharpsburg, PA, USA) and were projected onto the screen of MR-compatible LCD video goggles (VisuaStim XGA, Resonance Technology Inc.) worn by participants. The stimuli were presented at a resolution of 800 x 600 pixels and refreshed at 60 Hz). The three conditions (Fig. 1) were (i) no knowledge (NK), (ii) concealed knowledge (CK), and (iii) countermeasures (CM). Participants responded on a MR-compatible response box (Evoke Response Pad, Resonance Technology Inc.) using right-handed button presses with their index and middle finger (yes/no, respectively). They were instructed to respond as fast as possible without sacrificing accuracy. There were four runs for each condition and each run consisted of 36 trials where each digit was presented six times. Participants completed a total of 12 runs in the scanner.

All conditions included three types of items (single digits) as follows: (i) one "target" (16.7% of all stimuli), requiring participants to respond truthfully by pressing the "yes" button. This target number was given to participants before starting the fMRI session and it was included to ensure attention was paid in all tasks; (ii) one "probe" (16.7% of all stimuli), whose meaning and response instructions varied by condition as follows. For the NK condition, the probe was just another irrelevant number without any particular meaning for the participants, who simply pressed the "no" key to this item, indicating truthfully they did not know it. No information was given about the probe beforehand and so this control condition simulated the case of participants without concealed knowledge about the probe. In contrast, for the CK condition, participants were given the probe and they were instructed to keep this number secret from the experimenter by pressing the "no" button to pretend they did not know it. Finally, the CM condition was the same as the CK condition, with participants instructed to hide knowledge of the probe by pressing the "no" button. However, in this condition participants were also taught to perform a countermeasure on probe and irrelevants (CMpro, Fig 1). For the probe, the countermeasure consisted in focusing on superficial aspects of the stimulus, such as its physical features (e.g., font features, color, size, and so on) rather than its meaning. This way, the probe would become less meaningful to participants (iii) four "irrelevants" (66.7% of all stimuli), with no particular meaning for

participants, who responded truthfully by pressing the "no" button, indicating they did not know these numbers. In the CM condition, participants were also taught to perform a countermeasure on these irrelevants (CMirr, Fig 1) by focusing attention on memories associated with each of the stimuli (no specific examples were given in order to avoid limiting the kind of associations participants could spontaneously generate). This way, the irrelevants would become more meaningful to participants.

The reason for using two countermeasures is that the probe and the irrelevants were randomly intermixed in the task and so a single countermeasure could not be applied uniformly to all items because the saliency of the (infrequent) probe needed to be decreased whereas that of irrelevants needed to be increased for the countermeasure to be effective. Using only a probe countermeasure would increase the saliency of the probe, since this would be the only item to which the countermeasure is applied. Similarly, using only a countermeasure for irrelevants could also increase the saliency of the probe, as it would be the only item without a countermeasure (Meixner & Rosenfeld, 2010). Thus, two countermeasure components had to be designed, one for the probe and one for the irrelevants, so that participants would perform a countermeasure for all items. The countermeasures in the CM condition were based on those used by Uncapher and collaborators (Uncapher et al., 2015) and both were carried out by all participants during the CM condition, one on the probe and the other on the irrelevants. Note that in the CM condition too, participants were instructed to respond as fast as possible without sacrificing accuracy. This was done in order to minimize potential differences in the RTs between the CM and the CK conditions.

To ensure that the secret number had no meaning in the NK condition and that no systematic countermeasure strategies were applied during the CK condition, all participants completed the three conditions in the same order: NK, CK and CM, as in previous work (Ganis et al., 2011). Prior to the fMRI scan, participants underwent a practice session for the NK condition outside the scanner to familiarize them with the stimuli and responses. Instructions and practice for the CK and CM conditions were given before the actual runs while participants were in the scanner. The practice session consisted of 12 trials and was the same for each task but with different instructions.

To enhance the social component of the task so as to maximize participants' motivation,

before the CK condition participants were told that the main experimenter did not know the secret number, that she was not in the MRI control room, and that she would try to identify the secret number by analyzing their brain images on a computer in a different room. Thus, the job for participants during the CK and CM conditions was to hide their secret number by pressing the "no" button to pretend they did not know it while responding truthfully to all other numbers. In reality, the main experimenter knew the secret number and was in the MRI control room, and this was revealed during the debriefing session at the end of experiment. The debriefing session showed that all participants believed they were hiding their secret number from the experimenter, who was trying to discover it.

FIGURE1LINK

FMRI Data Acquisition

Whole brain imaging data were acquired using a Siemens Avanto whole body MRI scanner (1.5T, Siemens Medical System, Erlangen, Germany) with an 8-channels head coil. For each participant, functional images were acquired using a gradient-echo planar pulse T2*-weighted sequence with 31 axial slices in an interleaved ascending progression, parallel to the anterior-posterior commissural plane, TR = 2000 ms, FOV = 20 cm, 64 x 64 matrix and 3.125 x 3.125 x 4 mm resolution. During each run 116 functional volumes were acquired, for a total of 1392 volumes (116*12). A high resolution T1-weighted structural image was acquired, using a magnetization-prepared rapid gradient echo (MPRAGE) sequence with TR = 1900 ms, TE = 2.91 ms, FOV = 25 cm, FA = 8º, 176 sagittal slices, 256 x 256 matrix and 1 x 1 x 1 mm resolution for normalization to a template space. Finally, a T2-weighted structural image co-planar to the functional images was also obtained with 31 axial slices, TR = 7480 ms, TE = 94 ms, FA = 150º, 256 x 256 matrix and 0.8 x 0.8 x 4 mm resolution.

Preprocessing of fMRI Data

Brain imaging data were pre-processed and statistically analyzed using SPM8 (Statistical Parametric Mapping, Wellcome Trust Centre for Neuroimaging, London, UK).

The first four volumes of each run were not used in the analyses to ensure that T1 equilibrium was reached. For each participant, slice-time and motion correction were applied to the functional volumes, which were then coregistered to the co-planar T2-weighted images and finally with the T1-weighted structural images. T1 images were normalized to the Montreal Neurological Institute (MNI) template using SPM8's segmentation tool. The resulting transformation parameters obtained from the segmentation were applied to the functional images to spatially normalize them to MNI space (2 x 2 x 2 mm voxels for the whole brain univariate analyses, 3 x 3 x 3 mm voxels for the ROI and multivariate analyses). Finally, for the univariate analyses the normalized functional images were spatially smoothed using an 8 mm full-width at half-maximum (FWHM) Gaussian kernel.

Univariate fMRI Analyses

For the subject-level analyses, we applied voxel-wise univariate general linear models (GLM) on each participant's preprocessed functional data to obtain individual whole-brain estimates of brain responses to the stimuli presented during the NK, CK and CM conditions. For each condition and run, the GLM included one regressor for each of the 3 types of items (i.e., target, probe, and irrelevant) and six covariate motion parameters. Errors were modeled by one regressor of no interest. Onset delta functions were convolved with the canonical hemodynamic response function (HRF). Low-frequency noise was eliminated by high-pass filtering at 1/128 Hz.

For the whole-brain group analyses, statistical significance was first tested using an omnibus ANOVA (on data combined across runs in each condition) with Item (probe and irrelevants) and Condition (NK, CK, CM) as within-subject factors at $p < 0.001$ at the voxel level (uncorrected), with $p < 0.05$ (FWE-corrected) at the cluster level. Next, probe effect contrast images ("probe > irrelevants" and "irrelevants > probe") were generated to capture the difference in response between the probe and the irrelevants for each individual participant and for each condition. Follow-up analyses were conducted on these images to break down these effects.

To quantify the neural response in brain areas known to be engaged in this paradigm, we defined spherical ROIs (12 mm radius) around the centers of mass of the seven ROIs reported

in Ganis et al. (2011). In that study, the ROIs were identified by contrasting activation between the probe (the participant's date of birth) and the irrelevants (other unfamiliar dates) in a CIT paradigm with the same logic and temporal parameters as in the current study. It was critical to employ ROIs defined in an independent dataset in order to avoid overfitting (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). The coordinates of the centers of the seven ROIs were: 5, 21, 49 (medial/superior frontal gyrus/anterior cingulate, GFd), 1, -19, 34 (middle cingulate gyrus, GC), 45, 26, -6 (right inferior frontal gyrus/insula, RGFi), -38, 22, -8 (left inferior frontal gyrus/insula, LGFi), 52, -46, 42 (right inferior parietal lobule/supramarginal gyrus, RLPi), -60, -44, 35 (left inferior parietal lobule/supramarginal gyrus, LLPi), and 3, -3, 2 (thalamus, caudate nucleus, lenticular nucleus, Thal). Note that for simplicity in the rest of the paper we will refer to these ROIs by using only the first anatomical structure or its abbreviation, as done in the original study. An omnibus ANOVA with Item (probe and irrelevants), Condition (NK, CK, CM), and ROI (7 ROIs) as within-subject factors was conducted on the ROI activations. Follow-up ANOVAs and t-tests were carried out to break down these effects.

For all ANOVAs, *p* values involving factors with more than 2 levels were adjusted using the Greenhouse–Geisser correction for nonsphericity (Greenouse & Geisser, 1959).

Multivariate fMRI Analyses

Multivariate analyses were conducted to determine whether patterns of brain activation (multiple ROIs or multiple voxels) could reliably discriminate concealed and no concealed knowledge cases. The multivariate analyses were carried out on spatially normalized contrast images (probe minus irrelevants, averaging trials for each condition within a participant) without smoothing. Each feature was also normalized across cases by means of a z-score transformation (Hsu, Chang, & Lin, 2003). Classification analyses were carried out with the MATLAB implementation of LIBSVM (C. C. Chang & Lin, 2011). Since the number of features far exceeded the number of cases in the multi-voxel analyses, linear support vector machines (SVMs) instead of nonlinear ones were used (Hsu et al., 2003). All analyses reported here were conducted with default cost parameter c = 1. Exploratory analyses with lower and higher values of this parameter (range: 10^{-5} to 10^{5}) showed only small effects on

the results and so they will not be reported here. Additional exploratory analyses indicated that quadratic and Radial Basis Function kernels (RBFs) did not lead to better generalization than the linear kernel.

A key issue with multivariate analyses of fMRI data is the high-dimensionality of the datasets, usually requiring data reduction procedures before classification (Jin et al., 2009). To address this issue, in one set of analyses we used as features activation in the 7 ROIs defined in our previous study (Ganis et al., 2011), thus eliminating biases due to selecting features in the same dataset on which the classification is performed. Note that multivariate analyses in this context usually refer to "multi-voxel" analyses, where features are individual voxel activations (Tong & Pratte, 2012). However, they can also encompass "multi-ROI" analyses in which features are ROI average activations. Both multi-ROI and multivoxel analyses (voxels from the ROIs) were performed here.

A rigorous method to quantify the accuracy of the classification between two stimulus categories is to use signal detection theory (Green & Swets, 1966) and to measure the Area Under the (Receiver Operating Characteristic) Curve (AUC). The AUC quantifies the separation between two distributions independently of any specific classification threshold and has been recommended for the quantification of accuracy in deception research (National Research Council, 2003). An AUC equal to 1 indicates perfect classification accuracy whereas an AUC equal to 0.5 indicates classification at chance (see also the Validity Analysis section). A recent CIT study (Peth et al., 2015) used these methods to quantify the accuracy of the classification between the probe and irrelevants, and showed an Area Under the Curve (AUC) ranging between 0.50 and 0.87 using univariate ROI analyses, and higher classification accuracy (AUC ranging between 0.71 and 0.98) using activation in all grey matter voxels (over 10^5 features) as input to a linear classifier. We carried out a comparable analysis on our dataset.

To provide information about which voxels were important for the whole-brain classification, we also reconstructed the activation pattern from the classifier weight vector (for the CK/NK and CM/NK classifications separately, employing all data) using the forward modeling method described in Haufe and collaborators (Haufe et al., 2014).

In sum, the multivariate analyses were conducted on three types of data: multi-average

ROIs (7 features), multi-voxel ROIs (1069 features), multi-voxel whole brain (26452 features).

For the classification analyses, we used a one-pair-out cross-validation approach (Fig. 2) in which two cases per condition out of 40 were left out for testing, and training of the classifier was carried out on the remaining 36 cases (18 per condition). Since data in the different conditions were acquired in a within-subject manner, all cases for the left-out participants were excluded from the training set as well (e.g., if case CK for participant *Pi* was used during testing, then case NK for participant *Pi* was also excluded from the training set). Furthermore, cases from different conditions in the testing set were always from different participants so that testing was never carried out on conditions coming from the same participant. This was repeated for all possible 380 pairs of left-out cases (20x19). These analyses were repeated by allowing cases taking both members of the left-out pair from the same participant, but the differences were negligible, and so they will not be reported.

In some applied situations, one might build a classifier on a known set of NK and CK cases and then use it to classify new NK, CK, and CM cases. Thus, in these generalization analyses we trained a classifier to discriminate NK and CK cases, and compared the performance of the classifier on discriminating left-out NK vs CK cases and NK vs CM cases. The main difference with the previous procedure was that the left-out CK-NK and CM-NK cases were always from 4 different participants, which were not used during training to discriminate CK and NK cases. For these analyses, 10,000 random permutations of 4 participants (out of 20) were selected for testing, with training performed on the remaining participants (16 NK and 16 CK cases). The same analyses were repeated by removing the constraint that the left-out cases had to come from different participants, but the results were comparable and so they will not be reported here.

To determine how accurately a classifier discriminated between the different conditions, we used signal detection theory, as detailed in the following section (Peth et al., 2015).

FIGURE2LINK

Validity Analyses

The validity of behavioral and neural measures in discriminating pairs of conditions was calculated by generating receiver operating characteristic curves (National Research Council, 2003) using signal detection theory (Green & Swets, 1966). This approach provides more complete and precise information than simply calculating accuracy using hits and false alarms at a particular decision value threshold (criterion) that does not reflect directly the distance between the two distributions being discriminated. The key parameter estimated with these analyses is the area under the curve (AUC), which quantifies the separation between two distributions (for example, NK and CK cases) using information from the entire range of decision value thresholds (Green & Swets, 1966). An AUC equal to 1 indicates perfect classification accuracy whereas an AUC equal to 0.5 indicates classification at chance. In the present study, we carried out CK vs NK and CM vs NK classifications using the *activation probe effect* (probe minus irrelevants contrast estimate, for univariate fMRI data) and the decision value distribution for all possible left-out pairs (for multivariate fMRI data). Receiver Operating Characteristic curves for each possible pair of conditions were generated by calculating hits and false positives for criteria spanning the entire distribution of decision values. To determine whether a given AUC value was significantly different from chance (0.5), for univariate fMRI analyses we calculated the 95% confidence interval using parametric methods (Stanislaw & Todorov, 1999). For multivariate analyses, significance was determined using randomization methods (Good, 2005) to empirically estimate the null distribution of AUC values. To calculate the area under the curve (AUC) for the classification between conditions, we used the distribution of decision values resulting from this process (using 'svmpredict' in LIBSVM). The significance of the AUC was determined using a randomization approach as in previous work (Peth et al., 2015). Specifically, we estimated the null distribution of AUCs under the null hypothesis of no difference between conditions, by randomly shuffling the labels of the two conditions and by performing the classification procedure just described. This process was repeated 1000 times. The AUC calculated for the unshuffled data was considered significant at $p < 0.05$ if it was larger than 95% of the values in this null distribution (Fig. 2). The difference between the AUC in pairs of conditions was

tested for significance by estimating the null distribution of the difference by shuffling the data and by determining whether the unshuffled AUC difference was larger than 95% of the null difference distribution.

Results

Behavior

We conducted omnibus ANOVAs on the RTs end error rates (ERs), with item (probe, irrelevants) and condition (NK, CK, CM) as within-subject factors.

Results of the omnibus ANOVA for RTs showed a significant main effect of item and condition, as well as an interaction between the two (Table 1). Follow-up 2x2 ANOVAs and t-tests were carried out to unpack these results.

FIGURE3LINK

The ANOVA on the CK and NK conditions (Table 1 and Fig. 3a) showed a main effect of item type, and a trend for a main effect of condition. Importantly, there was a significant interaction between condition and item type. Follow-up t-tests showed that RTs to the probe were slower than to the irrelevants in the CK condition, $t(19) = 4.77$, $p < 0.0001$, but not in the NK condition, as found in previous studies using similar CIT paradigms (Ganis et al., 2011). These results provide a manipulation check for the concealed knowledge condition.

The ANOVA on the CM and NK conditions (Table 1 and Fig. 3a) showed main effects of item type and condition, as well as an interaction between condition and item type, indicating that the countermeasure manipulation had an effect. Follow-up t-tests revealed that RTs to the probe were slower than to the irrelevants in the CM condition, $t(19) = 3.15$, $p < .01$, but not in the NK condition. These results provide a manipulation check for the countermeasure condition.

Finally, the ANOVA on the CK and CM conditions (Table 1 and Fig. 3a) showed only a main effect of item type, as RTs were slower for the probe than the irrelevants. The mean RTs in the CM condition (554 ms) were numerically slower than those in the CK condition

(538 ms), but not significantly so. There was also a trend for the RTs to irrelevants (but not the probe) to be slower in the CM than in the CK condition, $t(19) = 1.77$, $p = 0.092$ (Fig 3a).

The same analyses were carried out for error rates (ERs, Fig. 3b), but the omnibus ANOVA showed no main effect of item ($F[1,19]=2.20$, $p = 0.16$, $\eta_p^2 = 0.10$), no main effect of condition ($F[2,38]=1.29$, $p = 0.29$, $\eta_p^2 = 0.06$), and no interaction between item and condition ($F[2,38]=2.46$, $p = 0.099$, $\eta_p^2 = 0.12$). Given the lack of significant effects in the omnibus ANOVA, no follow-up analyses were carried out on error rates.

FMRI: Univariate Whole-Brain Analyses

Fig. 4 shows the results of the whole-brain omnibus ANOVA, whereas Table 2 lists the peak coordinates and F-statistics. Of particular interest for the purpose of this study are the results of the interaction between the item and condition factors (Fig. 4, I x C), showing the regions in which the probe effect was modulated by condition type. These regions included the bilateral inferior frontal gyri and insula, the right middle frontal gyrus, the anterior cingulate cortex and medial frontal gyrus, and posterior parts of the middle and superior temporal gyri near the temporo-parietal junction. Follow-up analyses were conducted to break down this interaction.

FIGURE4LINK

Fig. 5 shows the activation probe effect maps (probe – irrelevant contrast) in the NK, CK and CM conditions, whereas Table 3 lists the peak coordinates. In the NK condition, this comparison yielded no significant activation clusters as expected, since participants did not have knowledge of the probe and so this was just another irrelevant. In contrast, in the CK condition, significantly stronger brain activation for the probe than the irrelevants was present in the middle/anterior cingulate cortex and medial frontal gyrus, the bilateral inferior frontal gyri and insula, the right precuneus, the right inferior parietal lobule, and the right caudate nucleus (Fig. 5 and Table 3). Finally, in the CM condition the activation was larger for the probe than the irrelevants in the right supramarginal gyrus whereas the opposite pattern was found in the medial orbitofrontal cortex.

Comparing directly the activation probe effect between the CK and NK conditions yielded essentially the same regions that were found in the CK condition (Table 4), confirming that the effects found in the CK condition were not due to stimulus peculiarities, as these would have been evident also in the NK condition. In contrast, comparing the activation probe effect between the CM and NK conditions and the CK and CM conditions yielded no significant differences.

FIGURE5LINK

FMRI: Univariate ROI Analyses

ROI analyses were used to complement the whole-brain analyses. We created 7 ROIs by defining 12 mm spheres around the center of mass of the ROIs reported in Ganis et al. (2011), as described earlier. These ROIs were used for both the univariate and the multivariate analyses. We conducted an omnibus ANOVA with item (probe, irrelevants), condition (NK, CK, CM), and ROI (LGFi, LLPi, GC, RGFi, RLPi, Thal, and GFd) as within-subject factors. This ANOVA showed a main effect of item since the probe elicited stronger activation than the irrelevants, an effect of condition indicating that activation was larger in the CK and CM than the NK condition, and an interaction between these two factors indicating that the probe effect varied across conditions (Table 5). Neither the main effect of the ROI factor nor the interactions involving this factor were significant. Follow-up ANOVAs were carried out to unpack these results.

The first analysis on the CK and NK conditions showed a main effect of item type and condition (Table 5 and Fig. 6), indicating that the probe elicited stronger activation than irrelevants and that the CK condition elicited stronger activation than the NK condition. The interaction between item type and condition indicated a larger probe effect in the CK than in the NK condition.

FIGURE6LINK

The second analysis comparing the CM and NK conditions also showed a main effect of item type and condition (Table 5 and Fig. 6), showing that the probe elicited stronger activation than irrelevants and that the CM condition elicited stronger activation than the NK condition. The interaction between item type and condition indicated that the probe effect was larger in the CM than in the NK condition.

Finally, the analysis on the CK and CM conditions showed a main effect of item type (Table 5 and Fig. 6), indicating that the probe elicited stronger activation than the irrelevants. The interaction between item type and condition indicated that the probe effect was larger in the CK than in the CM condition. Separate follow-up ANOVAs were conducted on the probe and irrelevants to further unpack this interaction. A main effect of condition revealed that activation to the probe was smaller in the CM than in the CK condition **(2.06 vs 3.22)**, $F(1,19)=5.19$, $p < 0.05$, $\eta_p^2= 0.21$. In contrast, although activation to the irrelevants was numerically larger in the CM than in the CK condition **(0.89 vs 0.71)**, there were no significant effects of condition in the analysis on the irrelevants. There was also an interaction between the ROI and Item factors, indicating that the probe effect varied by ROI. However, this interaction was not modulated by condition.

In sum, engaging in countermeasures reduced the size of the probe effect and this effect did not vary by ROI.

Effect of Block Order

To examine the potential role of repetition on the probe effect in the ROI analyses we carried out a repeated-measures ANOVA with Block (first two CK blocks vs last two CK blocks), Item Type (Probe vs Irrelevants), and ROI (7 ROIs). There was no main effect or block $F(1,19)=0.16$, $p = 0.69$, $\eta_p^2= 0.008$. Similarly, there were no interactions involving the block factor: Block x Item Type, $F(1,19)=0.03$, $p = 0.86$, $\eta_p^2= 0.002$; Block x ROI, $F(6,114)=1.62$, $p = 0.19$, $\eta_p^2= 0.08$; Block x Item Type x ROI, $F(6,114)=0.26$, $p = 0.31$, $\eta_p^2= 0.061$.

A similar ANOVA was conducted on the CM dataset. Results showed no main effect of block, $F(1,19)=0.44$, $p = 0.52$, $\eta_p^2= 0.022$, and no interactions involving the block factor: Block x Item Type, $F(1,19)=0.55$, $p = 0.47$, $\eta_p^2= 0.028$; Block x ROI, $F(6,114)=1.85$, $p =$

0.12, η_p^2= 0.09; Block x Item Type x ROI, $F(6,114)=0.32$, $p = 0.84$, η_p^2= 0.017. These results show that the probe effect did not decrease over time, indicating that the reduction of the probe effect in the CM task is not due to repetition effects.

FMRI Validity: Independent CK-NK and CM-NK Analyses

In these analyses, classifiers were trained independently to discriminate CK-NK and CM-NK cases and tested on discriminating left-out CK-NK and CM-NK cases, respectively.

Univariate ROI analyses. It was possible to discriminate above chance between CK and NK cases using the average activation from any of the seven ROIs (Table 6). The AUC varied among ROIs, and it was highest for the left inferior frontal gyrus (0.85), and lowest for the thalamus (0.73). In contrast, it was possible to discriminate CM and NK cases above chance only using activation in the right inferior frontal gyrus, the right inferior parietal lobule, and the left inferior frontal gyrus (0.74, 0.71, and 0.69, respectively). The countermeasures resulted in significantly smaller AUC values for the left inferior frontal gyrus, the middle cingulate gyrus, the thalamus, and the medial frontal gyrus (Hanley and McNeil, 1983). The same analyses carried out on the average of the seven ROIs (after score normalization) showed an AUC of 0.86 for discriminating CK versus NK cases, significantly larger than the AUC of 0.71 found for discriminating CM versus NK cases.

Multiaverage ROI analyses. In this analysis, average activation in each of the seven ROIs used in the univariate analyses was employed as a feature for a linear SVM classifier. The AUC for discriminating CK and NK cases was 0.85, significantly better than chance. The AUC for discriminating CM and NK cases was significantly smaller, 0.63, and not statistically different from chance (Table 7).

Multivoxel ROI analyses. In this analysis, activation in a total of 1069 voxels from the 7 ROIS (unsmoothed data) was used as input to a linear SVM classifier. The AUC for discriminating CK and NK cases was 0.83, whereas that for discriminating CM and NK cases was significantly smaller, 0.63, and not statistically different from chance (Table 7).

Multivoxel whole brain analyses. In this analysis, activation in a total of 26452 gray matter voxels (unsmoothed data) was used as input to a linear SVM classifier. The AUC for discriminating CK and NK cases was 0.80, whereas that for discriminating CM and NK cases was 0.79. Both AUCs were significantly better than chance, but they did not differ statistically from each other (Table 7).

Reconstruction of Activation Patterns. To estimate the contribution of each voxel to the CK-NK and CM-NK classifications the activation patterns were reconstructed using the classifier weight vector (Haufe et al., 2014). Fig. 4 shows that activation in a broad set of regions (hot colors) including the bilateral inferior frontal gyrus, the anterior and posterior cingulate and adjacent medial frontal cortex, and the bilateral parietal cortex, increases with prediction of knowledge cases. In contrast, in another set of regions (cold colors) including parts of the orbitofrontal cortex, the precuneus, and the cuneus, activation increases with prediction of no knowledge cases.

FIGURE7LINK

FMRI Validity: CK-CM Generalization Analyses

In these analyses a classifier was trained to discriminate NK and CK cases and then tested on discriminating left-out CK-NK and CM-NK cases (Table 7).

Multiaverage ROI analyses. The AUC for discriminating CK and NK cases was significantly larger than for discriminating CM and NK cases, 0.84 and 0.71, respectively. Both AUCs were significantly different from chance. At a fixed false alarm rate of 20%, the countermeasures reduced hit rates from 80% to 61%.

Multivoxel ROI analyses. The AUC for discriminating CK and NK cases was significantly larger than for discriminating CM and NK cases, 0.84 and 0.68, respectively.

Both AUCs were significantly different from chance. At a fixed false alarm rate of 20%, the countermeasures reduced hit rates from 75% to 37%.

Multivoxel whole-brain analyses. The AUCs for discriminating CK versus NK cases and CM versus NK cases were 0.79 and 0.73, respectively. Both AUCs were significantly different from chance, but they were not different from each other. At a fixed false alarm rate of 20%, the countermeasures reduced hit rates from 65% to 42%.

Discussion

This study found that activation in a set of prefrontal, parietal and subcortical regions differentiated between the probe and irrelevants in a concealed knowledge condition, but not in a matched no-knowledge condition, confirming and extending the results of previous fMRI (Ganis et al., 2011; Peth et al., 2015) and ERP-based (Jung, Kang, & Kim, 2013) studies. The activation probe effect (difference between the activation to probe and irrelevants) could be used to classify well above chance concealed and no-concealed knowledge cases. Critically, the mental countermeasures tested in this study reduced the size of the probe effect and decreased classification accuracy, at least with ROI analyses, extending the countermeasure findings of previous fMRI work (Ganis et al., 2011).

Activation Probe Effect

The pattern of brain activation for the probe effect was comparable to that found in previous fMRI CIT studies (Ganis et al., 2011; Peth et al., 2015).

First, the probe engaged the VLPFC (bilaterally) and the adjacent anterior insula (in the left hemisphere) more than the irrelevants. The substantial heterogeneity in the functional organization of these frontal regions, along both the rostro-caudal and the left-right dimensions, has made it difficult to determine their precise role in cognition (Levy & Wagner, 2011). One proposal is that the VLPFC and adjacent insula are involved in reflexive orienting of attention to behaviorally relevant changes in the environment. Indeed, these

regions, together with parts of the medial prefrontal cortex, have been considered a key component of a salience network (Seeley et al., 2007) and they have also been conceptualized as the frontal nodes in a largely right-lateralized ventral attentional network that includes the right inferior parietal cortex as well (Corbetta & Shulman, 2002). Another proposal has focused instead on the role of these frontal regions in motor inhibition processes (Aron, Robbins, & Poldrack, 2004; Swick, Ashley, & Turken, 2008), as they are consistently engaged for instance by Go/No-Go tasks. Meta-analytic approaches have shown that different subregions of the ventrolateral prefrontal and adjacent insular cortex tend to respond differently to tasks that tap into attentional reorienting and motor inhibition processes, though the segregation is not clear-cut (L. J. Chang, Yarkoni, Khaw, & Sanfey, 2013; Levy & Wagner, 2011). For instance, the pars opercularis of the right inferior frontal gyrus tends to be engaged by motor inhibition but not attentional reorienting tasks whereas the inferior frontal junction and the anterior insula tend to be engaged bilaterally (but with a right hemisphere bias) by both attentional reorienting and motor inhibition tasks (L. J. Chang et al., 2013; Levy & Wagner, 2011). In addition to attention reorienting and motor inhibition tasks, the VLPFC is usually engaged by other classes of tasks as well. Especially relevant for the CIT are the potential roles of the VLPFC in memory processes such as encoding and retrieval (e.g., Fletcher et al., 2002; Iidaka, Sadato, Yamada, & Yonekura, 2000) and in social cognitive processes such as action imitation (e.g., Levy & Wagner, 2011; Molnar-Szakacs, Iacoboni, Koski, & Mazziotta, 2005).

One or more of these processes could account for the concealed knowledge probe effect in VLPFC and the insula in CITs. Indeed, the probe is more salient than the irrelevants (since the probe is the only item associated with the episode to be concealed and it is presented infrequently) and lying to the probe is likely to require inhibiting a prepotent truthful response (Verschueure et al., 2011). Furthermore, the probe in the CK condition should engage retrieval processes more strongly than irrelevants because it is associated with a pre-experimental episode. Finally, it is also likely that the probe engages social cognitive processes more strongly than the irrelevants because deception only occurs for the probe and instructions usually mention that a judge would try to detect deception based on various deception cues.

The CIT literature has been mixed regarding the interpretation of the role of the VLPFC. This is in large part because CIT paradigms usually engage more than one of the processes just discussed and designing paradigms that isolate individual processes has proven to be very difficult. For instance, an fMRI study that attempted to eliminate response competition processes in the CIT interpreted VLPFC activation as reflecting memory-related processes (Gamer, Klimecki, Bauermann, Stoeter, & Vossel, 2012). Results from more recent work, however, have been interpreted as indicating that response competition processes are critical for VLPFC activation (Suchotzki, Verschuere, Peth, Crombez, & Gamer, 2015). A detailed discussion of why such discrepancies might exist goes beyond the scope of this paper, but the main point is that they are probably due to the difficulty in isolating individual processes in CIT paradigms.

Second, the probe engaged medial prefrontal cortical regions more than the irrelevants, including parts of the middle and anterior cingulate and the superior portions of the medial frontal gyrus. Portions of these medial prefrontal regions are activated by the same attentional and response inhibition tasks that engage the VLPFC and have been implicated, among other things, in monitoring the conflict between competing responses and related processes (Braver, Barch, Gray, Molfese, & Snyder, 2001; Rushworth, Walton, Kennerley, & Bannerman, 2004).

Third, portions of the right inferior parietal lobule, angular gyrus, and supramarginal gyrus were also recruited more strongly by the probe than the irrelevants. These regions overlap in large part with the right temporo-parietal junction (TPJ), which has been implicated in attentional reorienting in a number of domains (Corbetta & Shulman, 2002; Shomstein, 2012). More recently, the TPJ (bilaterally) has been suggested to be involved in contextual updating, that is in updating an internal model of the physical and social environment in order to revise expectations and responses after detecting a change in the environment (Geng & Vossel, 2013). Note that in previous CIT work the laterality of activation in the inferior parietal lobule has varied, with some studies reporting only activation in the right hemisphere (e.g., Gamer et al., 2012; Peth et al., 2015) and others in both hemispheres (e.g., Cui et al., 2014; Ganis et al., 2011), probably due to differences in stimuli and paradigms.

Fourth, portions of the right caudate nucleus also showed a probe effect. The caudate nucleus has been found also in other deception studies and its role has been generally attributed to response conflict detection (Kireev, Korotkov, Medvedeva, & Medvedev, 2013; Nunez, Casey, Egner, Hare, & Hirsch, 2005).

Finally, there was also a region in the right precuneus that has not been reported in previous CIT studies. Using the Neurosynth database (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011), the functional connectivity network defined by using this region as seed has the highest correlation with the functional connectivity networks associated with the terms "retrieval" and "memory retrieval" (r = 0.18). The potential memory retrieval role of this region would make sense because of the episodic memories associated with the probe. However, since this region was not found in previous CIT studies, we could not use it as an apriori region for the classification analyses.

Classification of CK and NK cases

With the exception of the precuneus, the pattern of activation for the CK – NK comparison is similar to that found in our previous study using the date of birth as probe (Ganis et al., 2011). Thus, the ROIs identified in that study were used to test the accuracy of single participant classification in the current study in order to avoid circularity in the analyses (Kriegeskorte et al., 2009). Results showed that the average probe effect in each of the 7 ROIs could be used to classify CK and NK cases well above chance. As found in the previous study, the ROIs with the highest accuracy were the left and right ventrolateral prefrontal cortex (AUC = 0.85 and 0.82, respectively), and the medial frontal cortex (AUC = 0.83). However, the accuracy rates were lower than in our previous study in which an AUC of 1 was found (Ganis et al., 2011). This may be due to a number of factors: i) a 1.5 T scanner was used in this study, which most likely produced data with a lower signal-to-noise ratio; ii) the stimuli used in this study (single digits) were less salient than those used in the previous study (one's date of birth); iii) the analyses were different, as this study used beta values rather than the number of significant voxels within ROIs, as input to the classifiers; iv) the population used in this study (University of Padova students and affiliates) was more varied in age and background than that used in the previous study (Harvard University

undergraduates), which may have increased the variance in the data.

The univariate ROI accuracy rates found in this study are similar to those found in the study by Peth and collaborators (Peth et al., 2015), where the AUC varied between .66 and .87, depending on the compared conditions. Similar accuracy rates were also found with the multivariate ROI analyses. However, the results of these two studies may not be directly comparable because our study did not classify directly the probe and irrelevants in various conditions (Peth et al., 2015), but rather it classified the probe effect (the difference between the probe and the irrelevants for each participant) between conditions (NK, CK, and CM). The advantage of classifying the probe effect across conditions is that it could help reduce potential individual differences in brain responses by subtracting activation to irrelevants from that to the probe in each individual separately.

Only a few studies have used whole-brain multivariate analyses on fMRI deception datasets, with mixed methods and results, as reviewed next. In one study, Davatzikos and collaborators applied SVM classification methods to an fMRI dataset collected during a modified CIT task in which participants concealed information about possessing a playing card (Davatzikos et al., 2005). Note that this study did not have a no-knowledge (NK) condition or group, and so the classification was conducted between the probe and irrelevants within a single group of participants with concealed knowledge (CK). As mentioned earlier, a key issue with multivariate analyses of fMRI data is the high-dimensionality of the datasets, usually requiring data reduction procedures before classification. In this study, fMRI data reduction before classification was carried out by averaging activation across nearby voxels (the data was downsampled to 16x16x16 mm “macrovoxels”, from the original 3.75x3.75x4 mm voxels). This resulted in 560 activation features that were then fed to a non-linear SVM with a Gaussian kernel. One-out analyses using individual average maps (as we did in the current study) led to an accuracy of about 88% (91% hit rate and 14% false alarm rate), corresponding to an AUC of about .93 (Grier, 1971).

A subsequent study (Jin et al., 2009) compared the classification accuracy with various feature selection methods on data collected in a previous study using a differentiation of deception paradigm (Kozel et al., 2005). The authors reported that using a leave-one-out cross-validation scheme on data from all gray matter voxels (65,166 features) without any

data reduction led to a poor SVM classification rate of less than 60% (AUC < .65). Furthermore, resampling the data to a 16x16x16 mm grid as done by Davatzikos and collaborators did not improve the results. Higher accuracy rates were found instead with a number of data reduction methods such as Fisher criterion score, and so these authors recommended performing data reduction before classification (Jin et al., 2009).

A recent study used multivariate analyses with linear SVMs on data from a CIT paradigm and reported a maximum AUC of .98 for classifying probes in a concealed knowledge group versus irrelevants in a no-knowledge group by employing 125,570 grey matter voxels (whole brain data) without any data reduction (Peth et al., 2015). However, classification accuracy using whole brain data was not consistently high because the AUC for classifying the probe and irrelevants in a guilty intention group was only 0.71, numerically lower than with univariate ROI analyses, and not significantly different from chance. Furthermore, it is not clear that whole brain analyses were significantly more accurate than the corresponding univariate or multivariate ROI analyses. For example, the AUC for univariate analyses using combined ROIs was .87 (Guilty Action probe versus Guilty Intention irrelevants), which is not significantly higher than the AUC of .90 found for the corresponding whole-brain multi-voxel analysis (Peth et al., 2015). This variability in the literature suggests that idiosyncracies in relatively small datasets may play an important role in the results.

Overall, these accuracies are comparable to those found in electrophysiological and psychophysiological studies using the CIT. Indeed, a recent meta-analysis of the CIT using skin conductance and event-related potential measures (P300 amplitude, specifically) reported effect sizes (d) of about 1.55 and 1.89, respectively (Meijer, Selle, Elber, & Ben-Shakhar, 2014), corresponding to AUCs of .86 and .91, respectively (Rice & Harris, 2005). Note however, that a meaningful comparison between different techniques is difficult because of the differences in paradigm details. For example, typically event-related potential studies present the stimuli at a much faster rate than fMRI studies do.

In the current study, whole brain multivoxel analyses did not improve classification accuracy of CK versus NK cases compared to multivoxel ROI analyses (Table 6). A potential explanation for this is that the 7 ROIs already included voxels that maximally discriminated

between the CK and NK conditions (given that they were defined in a previous study contrasting these two conditions) and so including voxels outside these ROIs may have added more noise than additional information useful for the classification, resulting in a slightly degraded performance.

Effect of Mental Countermeasures

The VLPFC, insula, and medial prefrontal regions discussed earlier have high activation base rates, which means that they are engaged by many different tasks (Christ, Van Essen, Watson, Brubaker, & McDermott, 2009; Yarkoni et al., 2011). An important consequence of this is that CIT methods that rely on activation in these regions are likely to be vulnerable to mental countermeasures because a number of cognitive processes unrelated to concealed information can be engaged to modulate neural responses in these regions (Ganis, 2018a). Indeed, the physical countermeasures used in our previous study reduced the size of the probe effect and classification accuracy (Ganis et al., 2011).

The current fMRI study shows that mental countermeasures are also effective, at least when using ROIs with high base rates. Mental countermeasures reduced the size of the probe effect across the 7 ROIs, relative to the CK condition, mostly by decreasing activation elicited by the probe. Activation to irrelevants increased as a result of the countermeasure, but not significantly so. This may be due in part to the nature of saliency itself, which is inversely related to frequency of occurrence, and thus it is possible that trying to make all irrelevants more salient at the same time was much less effective than trying to increase the saliency of a single irrelevant. The results are consistent with those found in our previous study in which countermeasures were applied to a subset of the irrelevants (Ganis et al., 2011).

Classification accuracy was significantly lower for CM-NK than CK-NK cases in both univariate and multivariate ROI analyses, indicating that for these analyses countermeasures were effective. For the univariate analyses, the AUC decreased from .86 to .71 when combining ROIs. For the multiaverage ROI analyses, accuracy decreased from .85 to .63, and for the multivoxel ROI analyses it decreased from .83 to .63 (Table 6). However, the mental countermeasures had virtually no effect on the AUC based on whole-brain multivoxel

analyses, 0.8 and 0.79, for CK-NK and CM-NK cases, respectively. This is consistent with the idea that regions outside the 7 ROIs that are useful for the classification may be less vulnerable to this type of countermeasures.

Relative to multivoxel ROI analyses, multivoxel whole-brain analyses improved the AUC for CM-NK classification, from .63 to .79. This result is in contrast with what was found for CK-NK classification, and it may be explained by assuming that voxels that maximally discriminated between CM and NK cases were not included in the 7ROIs, since these were defined by the CK-NK contrast, and so adding other regions improved classification accuracy for CM-NK cases.

To test this idea, we ran a batch of multivoxel analyses that included the 7 original ROIs and gradually added more and more spherical ROIs (up to 200, radius = 12mm) centered at random locations. Results showed that, as the number of additional ROIs grew to 200, average CK-NK classification accuracy slightly declined (gradually going from .83 to .80), whereas average CM-NK classification accuracy gradually increased from .63 to .79. This suggests that the information useful for the CK-NK classification is already available within the original ROIs, whereas additional information about the CM-NK classification is distributed across voxels outside the original ROIs.

This interpretation was also confirmed by a multivoxel analysis carried out on all voxels not included in the 7 ROIs (radius = 20 mm). Results showed that, compared to using only the 7 ROIs, CK-NK classification accuracy decreased from .83 to .74 whereas CM-NK classification accuracy increased from .63 to .70. In other words, collectively, there was more information useful for the CM-NK classification in voxels outside than inside the 7 ROIs, whereas the reverse was true for the CK-NK classification.

The reconstructed activations from the whole brain classification (Fig. 4) show a distributed and complex pattern, as found in previous work (Peth et al., 2015). Activity increases in some regions, such as the bilateral inferior frontal gyrus and the cingulate gyrus, predicted concealed knowledge in both the CK and CM conditions. In contrast, activity increases in other regions, such as the orbitofrontal cortex, the precuneus, and the cuneus, predicted the absence of concealed knowledge only in the CM condition.

The CK-CM generalization analyses showed a similar pattern of effects (Table 6), with

countermeasures significantly decreasing accuracy in the multiaverage and multivoxel ROI analyses, but not in the whole brain multivoxel analyses.

Although the classification accuracy of CK-NK cases with whole-brain multivoxel analyses was rather low in this study (AUC = 0.8), the findings suggest that these analyses may be more robust to mental countermeasures than ROI ones, possibly because they include activation from regions with lower base rates that is not as easily modulated by the processes engaged by the countermeasures.

Physical or Mental Countermeasures?

The distinction between physical and mental countermeasures is not clear-cut, as mentioned in the introduction. Typically, the distinction is made based on whether the countermeasure requires physical actions (e.g., biting one's tongue) or not (e.g., reactivating certain memories). However, some of the brain mechanisms by which a countermeasure works may be very similar in the two cases. For instance, physical countermeasures that result in changes in autonomic parameters such as skin conductance could also elicit brain activation that interferes with the salience network engaged by the CIT. Indeed, activation in the right insular region engaged by the CIT has been shown to correlate with skin conductance changes acquired simultaneously during the MRI session (Gamer, Bauermann, Stoeter, & Vossel, 2007). Thus, a physical countermeasure that produces a change in skin conductance (e.g., self-inflicting pain) might interfere with the accuracy of fMRI-based CITs. These kinds of autonomic changes could be indirectly produced also by mental countermeasures and contribute to the effectiveness of the countermeasures. As a more general point, many nodes of the salience network engaged by the CIT have high base rates, that is, they can also be engaged by many other cognitive and affective tasks (Ganis, 2018a, 2018b); therefore, it is likely that activation in these nodes can be altered by many types of countermeasures, regardless of whether they are classified as physical or mental.

Limitations

Although this study provides new evidence on the effect of mental countermeasures in

fMRI-based CIT paradigms, some limitations should be mentioned. First, the paradigm and stimuli were intentionally minimalistic in order to reduce potential perceptual difference between stimuli that can be problematic for the interpretation of CIT results (Ganis et al., 2016). This approach was justified here because: i) employing numbers as stimuli could be used in various forensic situations, for example, where a certain sum of money was stolen, where a number was part of the crime scene (e.g., the number of the apartment where a crime took place), or where somebody may be in possession of certain secret information (e.g., a bank account number); in that respect, numbers are not different from other types of crime relevant items and they may even be more generalizable than using specific objects (e.g., a watch), ii) the main goal of the study was to determine whether the accuracy of 3S CIT paradigms is affected by mental countermeasures, regardless of absolute accuracy rates in the field, and iii) even elaborate mock crime scenarios are generally far from ecologically valid situations because participants still know that the entire situation is fictitious, they are just following instructions (and so, the probe is not a crime item, nothing unlawful is done by "stealing" it), and the items employed in the scenarios don't usually have much of an intrinsic value to the participants.

Second, the tasks were administered in a within-participant manner and in a fixed order, as in our previous study (Ganis et al 2011). To eliminate the effect of potential within-subject correlations during testing, in the classification analyses we used test cases from participants that had been excluded from the training set. For example, if the case NK for a participant was used for testing, the CK and CM cases for that participant were also excluded from training. The analyses were repeated by removing this constraint as well, but the results were comparable. The presence of within-subject correlations between conditions during training may affect the absolute accuracy, either increase it or decrease it, depending on whether it makes classification easier or more difficult.

Our previous CIT study on physical countermeasures provided initial evidence that the effect of the countermeasures in these paradigms are not due to order artifacts as there were no systematic order-related differences among the blocks in the tasks. Similarly, in this study, we compared the size of the probe effect during the first two and last two blocks in the CK condition and found no significant effects of block.

Third, although the number of cases (20 per condition) was comparable to that used in previous studies, this is still a relatively small number of cases to assess classification accuracy, especially given the large number of features used in the multivariate classification analyses. Larger CIT datasets will be needed in the future to further explore the potential of machine learning methods in this domain.

Conclusions

These results confirm that a probe (familiar item) is associated with an increase in activation in an attentional fronto-parietal network, relative to irrelevants (unfamiliar items). Using a CIT paradigm, activation in this network can be used to determine if an individual is familiar with the probe. The accuracy for this classification was comparable between ROI and whole-brain analyses. Critically, classification accuracy was reduced by mental countermeasures that artificially manipulated attentional allocation to the probe and irrelevants. However, the effect of mental countermeasures was only statistically significant for the ROI analyses. For the whole brain analyses, classification accuracy was relatively low (AUC = 0.8), but it was not significantly reduced by the countermeasures, suggesting that whole-brain multivariate analyses may hold some potential to reduce the effect of mental countermeasures. Overall, these findings remind us that the effect of physical and mental countermeasures needs to be taken into account before these methods can be applied in the field. Finally, further progress in the field may take place in at least three ways: i) by devising methods to detect the use of mental countermeasures, ii) by devising counter-countermeasures, or iii) by adapting to fMRI and testing paradigms that have shown resistance to countermeasures with ERP methods (Bowman et al., 2013; Rosenfeld et al., 2008).

Figure legends

Figure 1. Diagram of the experimental design. In the no knowledge condition (NK), participants told the truth on all trials, responding "Yes" to the target (shown in red in the figure) and "No" to all other stimuli. Note that participants had no knowledge about the probe in this condition. Participants learned about the probe (secret number, shown in green in the figure) just before the concealed knowledge condition (CK). In this condition, participants used the same pattern of responses, but they lied about knowing the probe. Finally, the countermeasure condition (CM) was the same as the concealed knowledge condition, but participants also carried out countermeasures for irrelevants (CMirr) and for the probe (CMpro).

Figure 2. Diagram illustrating the logic of the multivariate classification analyses. In the data matrices, each case is a row and each column is a feature (ROI or voxel). For example, row 1 of C1 contains the data for participant 1 tested in condition 1. In the main analysis, a classifier is trained on N-2 cases per condition and then tested on left-out cases, with instances for the different conditions always from different participants. For example, if the left-out case for C1 is from participant N-1, then the left-out case for C2 is from participant N. The process is repeated for all possible left-out testing pairs, allowing the computation of the area under the receiver operating characteristic curve (AUC) of the dataset of interest (blue curve). Significance testing is carried out by computing the null distribution of AUCs calculated as above, but on the dataset with randomly shuffled labels (1,000 random shuffles, black curves). An AUC is considered significantly different from chance at $p = 0.05$ if its value is larger than 95% of the AUCs in the null distribution. Testing the significance between AUCs is carried out using a similar logic by computing the null distribution of the difference between AUCs. The generalization analyses follow the same logic, but training is carried out on N-4 cases in order to leave enough cases for testing, and the process is repeated 10,000 times

Figure 3. Behavioral results for the probe and irrelevants in the NK, CK, and CM conditions.

a) Response times, and b) error rates. Error bars represent the standard error of the mean.

Figure 4. Activation maps showing the results of the omnibus ANOVA with Item (Probe and Irrelevants) and condition (NK, CK, CM) as factors. The first column shows the main effect of Item. The second column shows the main effect of condition. The last column shows the interaction between item and condition. Warmer colors represent higher *F* values.

Figure 5. Activation maps for the probe effect (contrast probe minus irrelevants) in the CK (left column) and CM (right column) conditions. There was no activation in the NK condition, as expected because in this condition participants have no knowledge about the probe. Warmer colors represent higher *t* values.

Figure 6. Activation difference (AU, arbitrary units) between the probe and irrelevants in the 7 ROIs (Ganis et al., 2011) in the NK, CK, and CM conditions. Error bars represent the standard error of the mean.

Figure 7. Reconstructed activation patterns (Haufe et al., 2014) for CK/NK and CM/NLK classification overlaid on the icbm_avg_152T1 template for 5 horizontal slices (z = -20, -6, 8, 22, 36, from left to right). Voxels in which activation increases with prediction of knowledge cases are indicated in hot colors, whereas voxels in which activation increases with prediction of no knowledge cases are indicated in cold colors. Only voxels with weights larger than 0.5 in absolute value are shown.

References

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences, 8*(4), 170-177. doi:10.1016/j.tics.2004.02.010

Ben-Shakhar, G. (2012). Current research and potential applications of the concealed information test: an overview. *Frontiers in Psychology, 3*, 342. doi:10.3389/fpsyg.2012.00342

Bowman, H., Filetti, M., Janssen, D., Su, L., Alsufyani, A., & Wyble, B. (2013). Subliminal salience search illustrated: EEG identity and deception detection on the fringe of awareness. *PloS One, 8*(1), e54258. doi:10.1371/journal.pone.0054258

Braver, T. S., Barch, D. M., Gray, J. R., Molfese, D. L., & Snyder, A. (2001). Anterior cingulate cortex and response conflict: effects of frequency, inhibition and errors. *Cerebral Cortex, 11*(9), 825-836.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol., 2*, 1-27.

Chang, L. J., Yarkoni, T., Khaw, M. W., & Sanfey, A. G. (2013). Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. *Cerebral Cortex, 23*(3), 739-749. doi:10.1093/cercor/bhs065

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex, 19*(7), 1557-1566. doi:10.1093/cercor/bhn189

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews: Neuroscience, 3*(3), 201-215. doi:10.1038/nrn755

Cui, Q., Vanman, E. J., Wei, D., Yang, W., Jia, L., & Zhang, Q. (2014). Detection of deception based on fMRI activation patterns underlying the production of a deceptive response and receiving feedback about the success of the deception after a mock murder crime. *Social Cognitive and Affective Neuroscience, 9*(10), 1472-1480. doi:10.1093/scan/nst134

Dale, A. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping, 8*, 109-114.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., . . .

Langleben, D. D. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage, 28*(3), 663-668. doi:10.1016/j.neuroimage.2005.08.009

Fletcher, P. C., Palomero-Gallagher, N., Zafiris, O., Fink, G. R., Tyler, L. K., & Zilles, K. (2002). The influence of explicit instructions and stimulus material on lateral frontal responses to an encoding task. *Neuroimage, 17*(2), 780-791.

Gamer, M., Bauermann, T., Stoeter, P., & Vossel, G. (2007). Covariations among fMRI, skin conductance, and behavioral data during processing of concealed information. *Human Brain Mapping, 28*(12), 1287-1301. doi:10.1002/hbm.20343

Gamer, M., Klimecki, O., Bauermann, T., Stoeter, P., & Vossel, G. (2012). fMRI-activation patterns in the detection of concealed information rely on memory-related effects. *Social Cognitive and Affective Neuroscience, 7*(5), 506-515. doi:10.1093/scan/nsp005

Ganis, G. (2018a). Detecting Deception and Concealed Information with Neuroimaging. In J. P. Rosenfeld (Ed.), *Detecting Concealed Information and Deception*: Academic Press.

Ganis, G. (2018b). Lying and neuroscience. In J. Meibauer (Ed.), *The Oxford handbook of lying* (pp. 455-468): Oxford University Press.

Ganis, G., Bridges, D., Hsu, C. W., & Schendan, H. E. (2016). Is anterior N2 enhancement a reliable electrophysiological index of concealed information? *Neuroimage, 143*, 152-165. doi:10.1016/j.neuroimage.2016.08.042

Ganis, G., Rosenfeld, J. P., Meixner, J., Kievit, R. A., & Schendan, H. E. (2011). Lying in the scanner: covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *Neuroimage, 55*(1), 312-319. doi:10.1016/j.neuroimage.2010.11.025

Geng, J. J., & Vossel, S. (2013). Re-evaluating the role of TPJ in attentional control: contextual updating? *Neuroscience and Biobehavioral Reviews, 37*(10 Pt 2), 2608-2620. doi:10.1016/j.neubiorev.2013.08.010

Good, P. L. (2005). *Resampling methods: 3d Edition*: Birkhauser.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: John Wiley and Sons Inc.

Greenouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: computing formulas. *Psychological Bulletin, 75*(6), 424-429.

Haufe, S., Meinecke, F., Gorgen, K., Dahne, S., Haynes, J. D., Blankertz, B., & Biessmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage, 87*, 96-110. doi:10.1016/j.neuroimage.2013.10.067

Honts, C. R., Devitt, M. K., Winbush, M., & Kircher, J. C. (1996). Mental and physical countermeasures reduce the accuracy of the concealed knowledge test. *Psychophysiology, 33*(1), 84-92.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). *A Practical Guide to Support Vector Classification*. Technical Report. Department of Computer Science. National Taiwan University. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Iidaka, T., Sadato, N., Yamada, H., & Yonekura, Y. (2000). Functional asymmetry of human prefrontal cortex in verbal and non-verbal episodic memory as revealed by fMRI. *Brain Research: Cognitive Brain Research, 9*(1), 73-83.

Jin, B., Strasburger, A., Laken, S. J., Kozel, F. A., Johnson, K. A., George, M. S., & Lu, X. (2009). Feature selection for fMRI-based deception detection. *BMC Bioinformatics, 10 Suppl 9*, S15. doi:10.1186/1471-2105-10-S9-S15

Jung, E. K., Kang, K. Y., & Kim, Y. Y. (2013). Frontoparietal activity during deceptive responses in the P300-based guilty knowledge test: an sLORETA study. *Neuroimage, 78*, 305-315. doi:10.1016/j.neuroimage.2013.04.027

Kireev, M., Korotkov, A., Medvedeva, N., & Medvedev, S. (2013). Possible role of an error detection mechanism in brain processing of deception: PET-fMRI study. *International Journal of Psychophysiology, 90*(3), 291-299. doi:10.1016/j.ijpsycho.2013.09.005

Kozel, F. A., Johnson, K. A., Mu, Q., Grenesko, E. L., Laken, S. J., & George, M. S. (2005). Detecting deception using functional magnetic resonance imaging. *Biological Psychiatry, 58*(8), 605-613. doi:10.1016/j.biopsych.2005.07.040

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience, 12*(5), 535-540. doi:10.1038/nn.2303

Levy, B. J., & Wagner, A. D. (2011). Cognitive control and right ventrolateral prefrontal cortex: reflexive reorienting, motor inhibition, and action updating. *Annals of the New York Academy of Sciences, 1224*, 40-62. doi:10.1111/j.1749-6632.2011.05958.x

Lykken, D. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*, 385-388.

Matsuda, I., Nittono, H., & Allen, J. J. (2012). The current and future status of the concealed information test for field use. *Frontiers in Psychology, 3*, 532. doi:10.3389/fpsyg.2012.00532

Meijer, E. H., Selle, N. K., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: a meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology, 51*(9), 879-904. doi:10.1111/psyp.12239

Meixner, J. B., & Rosenfeld, J. P. (2010). Countermeasure mechanisms in a P300-based concealed information test. *Psychophysiology, 47*(1), 57-65. doi:10.1111/j.1469-8986.2009.00883.x

Molnar-Szakacs, I., Iacoboni, M., Koski, L., & Mazziotta, J. C. (2005). Functional segregation within pars opercularis of the inferior frontal gyrus: evidence from fMRI studies of imitation and action observation. *Cerebral Cortex, 15*(7), 986-994. doi:10.1093/cercor/bhh199

National Research Council. (2003). *The Polygraph and Lie Detection. Committee to Review the Scientific Evidence on the Polygraph.* Washington, DC: The National Academic Press.

Nunez, J. M., Casey, B. J., Egner, T., Hare, T., & Hirsch, J. (2005). Intentional false responding shares neural substrates with response conflict and cognitive control. *Neuroimage, 25*(1), 267-277. doi:10.1016/j.neuroimage.2004.10.041

Peth, J., Sommer, T., Hebart, M. N., Vossel, G., Buchel, C., & Gamer, M. (2015). Memory detection using fMRI - does the encoding context matter? *Neuroimage, 113*, 164-174. doi:10.1016/j.neuroimage.2015.03.051

Rice, M. E., & Harris, G. T. (2005). Comparing Effect Sizes in Follow-Up Studies:ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 615-620.

Rissman, J., Greely, H. T., & Wagner, A. D. (2010). Detecting individual memories through the neural decoding of memory states and past experience. *Proceedings of the National Academy of Sciences of the United States of America, 107*(21), 9849-9854. doi:10.1073/pnas.1001028107

Rosenfeld, J. P., Ben-Shakhar, G., & Ganis, G. (2012). Detection of concealed stored memories with psychophysiological and neuroimaging methods. In L. Nadel & W. P. Sinnott-Armstrong (Eds.), *Memory and the Law*. New York: Oxford University Press.

Rosenfeld, J. P., Labkovsky, E., Winograd, M., Lui, M. A., Vandenboom, C., & Chedid, E. (2008). The Complex Trial Protocol (CTP): a new, countermeasure-resistant, accurate, P300-based method for detection of concealed information.

Psychophysiology, 45(6), 906-919. doi:10.1111/j.1469-8986.2008.00708.x

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology, 41*(2), 205-219.

Rushworth, M. F., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences, 8*(9), 410-417. doi:10.1016/j.tics.2004.07.009

Shomstein, S. (2012). Cognitive functions of the posterior parietal cortex: top-down and bottom-up attentional control. *Frontiers in Integrative Neuroscience, 6*, 38. doi:10.3389/fnint.2012.00038

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149.

Suchotzki, K., Verschuere, B., Peth, J., Crombez, G., & Gamer, M. (2015). Manipulating item proportion and deception reveals crucial dissociation between behavioral, autonomic, and neural indices of concealed information. *Human Brain Mapping, 36*(2), 427-439. doi:10.1002/hbm.22637

Swick, D., Ashley, V., & Turken, A. U. (2008). Left inferior frontal gyrus is critical for response inhibition. *BMC Neuroscience, 9*, 102. doi:10.1186/1471-2202-9-102

Tong, F., & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual Review of Psychology, 63*, 483-509. doi:10.1146/annurev-psych-120710-100412

Uncapher, M. R., Boyd-Meredith, J. T., Chow, T. E., Rissman, J., & Wagner, A. D. (2015). Goal-Directed Modulation of Neural Memory Patterns: Implications for fMRI-Based Memory Detection. *Journal of Neuroscience, 35*(22), 8531-8545. doi:10.1523/JNEUROSCI.5145-14.2015

Verschueure, B., Ben-Shakhar, G., & Meijer, E. (Eds.). (2011). *Memory Detection*. Cambridge, UK: Cambridge University Press.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods, 8*(8), 665-670. doi:10.1038/nmeth.1635

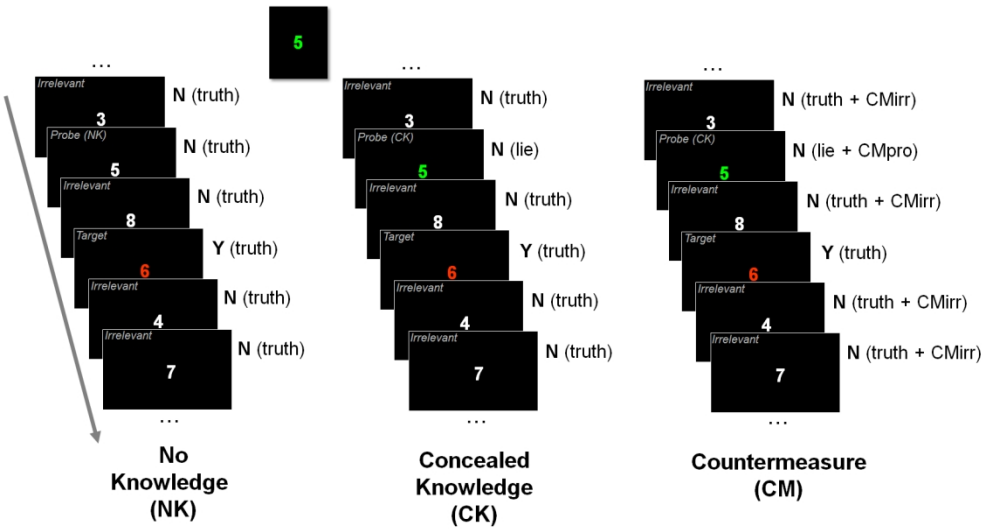

Figure 1. Diagram of the experimental design. In the no knowledge condition (NK), participants told the truth on all trials, responding "Yes" to the target (shown in red in the figure) and "No" to all other stimuli. Note that participants had no knowledge about the probe in this condition. Participants learned about the probe (secret number, shown in green in the figure) just before the concealed knowledge condition (CK). In this condition, participants used the same pattern of responses, but they lied about knowing the probe. Finally, the countermeasure condition (CM) was the same as the concealed knowledge condition, but participants also carried out countermeasures for irrelevants (CMirr) and for the probe (CMpro).

471x252mm (96 x 96 DPI)

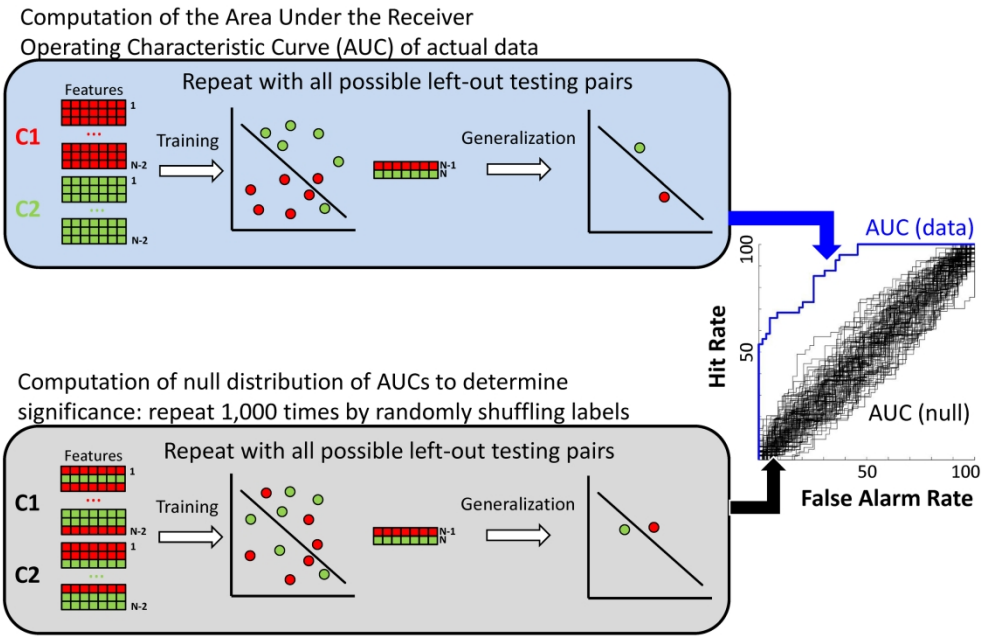

Figure 2. Diagram illustrating the logic of the multivariate classification analyses. In the data matrices, each case is a row and each column is a feature (ROI or voxel). For example, row 1 of C1 contains the data for participant 1 tested in condition 1. In the main analysis, a classifier is trained on N-2 cases per condition and then tested on left-out cases, with instances for the different conditions always from different participants. For example, if the left-out case for C1 is from participant N-1, then the left-out case for C2 is from participant N. The process is repeated for all possible left-out testing pairs, allowing the computation of the area under the receiver operating characteristic curve (AUC) of the dataset of interest (top blue curve). Significance testing is carried out by computing the null distribution of AUCs calculated as above, but on the dataset with randomly shuffled labels (1,000 random shuffles, cloud of black curves). An AUC is considered significantly different from chance at p = 0.05 if its value is larger than 95% of the AUCs in the null distribution. Testing the significance between AUCs is carried out using a similar logic by computing the null distribution of the difference between AUCs. The generalization analyses follow the same logic, but training is carried out on N-4 cases in order to leave enough cases for testing, and the process is repeated 10,000 times

254x190mm (300 x 300 DPI)

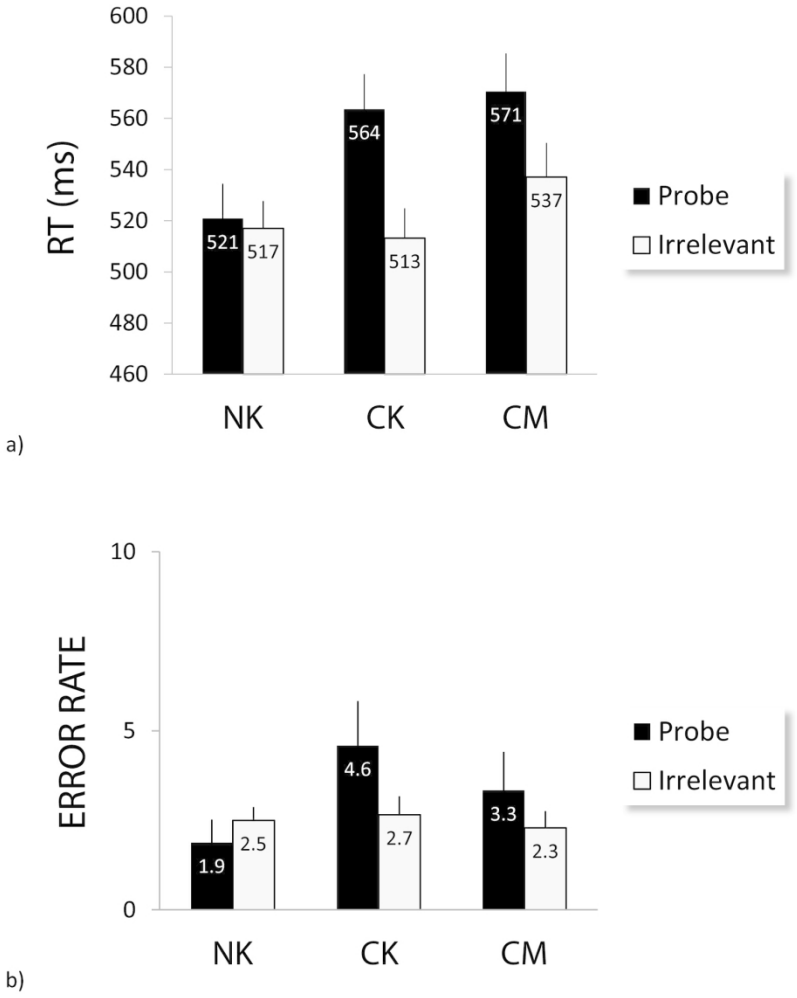

Figure 3. Behavioral results for the probe and irrelevants in the NK, CK, and CM conditions. a) Response times, and b) error rates. Error bars represent the standard error of the mean.

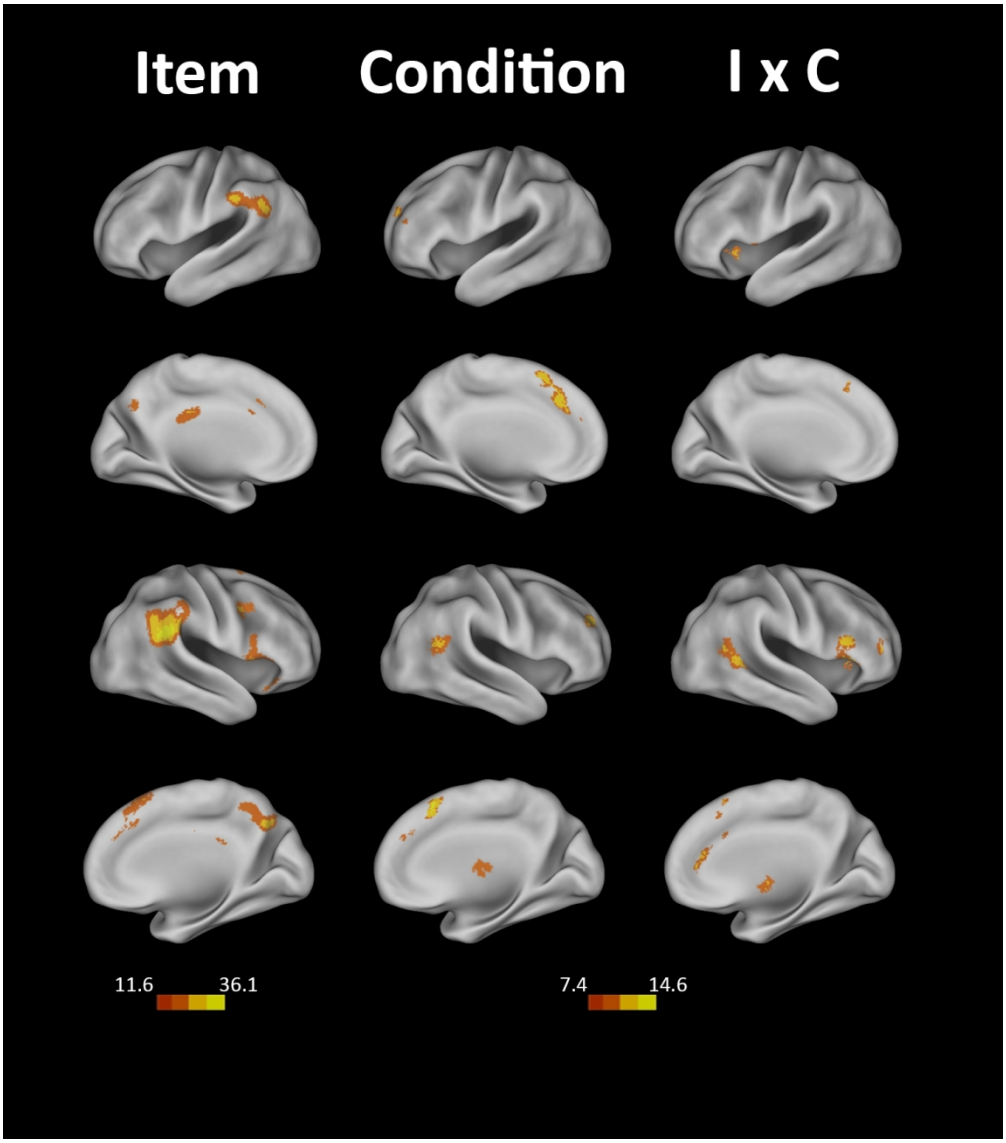

Figure 4. Activation maps showing the results of the omnibus ANOVA with Item (Probe and Irrelevants) and condition (NK, CK, CM) as factors. The first column shows the main effect of Item. The second column shows the main effect of condition. The last column shows the interaction between item and condition. Warmer colors represent higher F values.

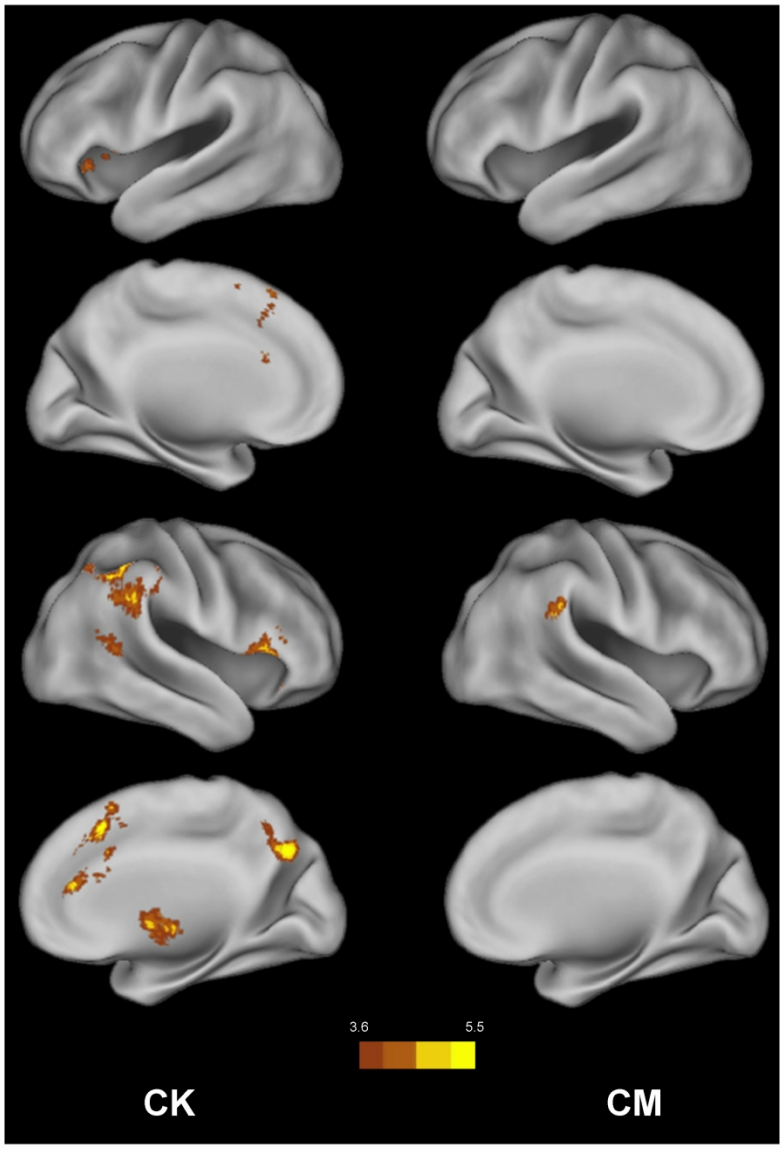

Figure 5. Activation maps for the probe effect (contrast probe minus irrelevants) in the CK (left column) and CM (right column) conditions. There was no activation in the NK condition, as expected because in this condition participants have no knowledge about the probe. Warmer colors represent higher t values.

215x279mm (300 x 300 DPI)

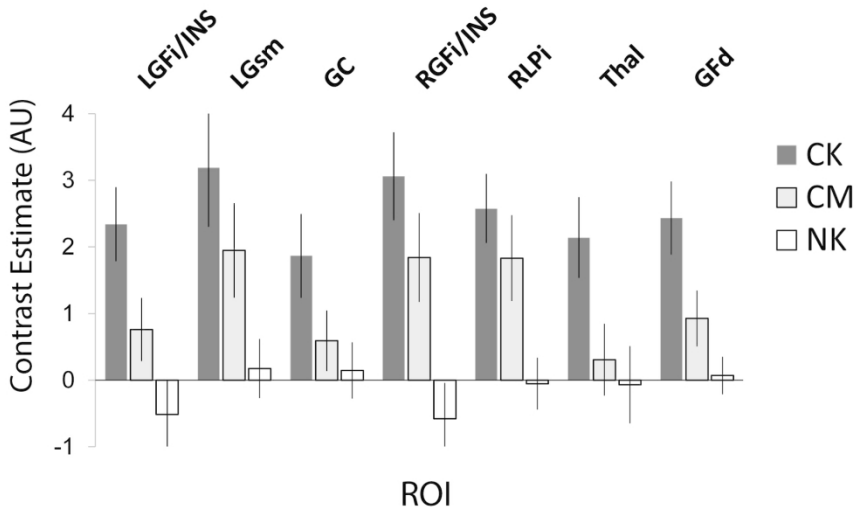

Figure 6. Activation difference (AU, arbitrary units) between the probe and irrelevants in the 7 ROIs (Ganis et al., 2011) in the NK, CK, and CM conditions. Error bars represent the standard error of the mean.

Table 1. ANOVAs comparing response times (RTs) for the probe and the irrelevants across conditions (omnibus, CK-NK-CM; condition pairs, CK-NK, CM-NK, CK-CM).

	Omnibus			Pairwise								
	CK / NK / CM			CK / NK			CM / NK			CK / CM		
Source	F	p	η_p^2	F	p	η_p^2	F	p	η_p^2	F	p	η_p^2
Item	17.97	<0.001	0.49	12.65	<0.005	0.4	8.09	<0.01	0.30	23.28	<0.001	0.55
Condition	3.55	<0.05	0.16	4.16	0.055	0.18	5.59	<0.05	0.23	1.14	0.30	0.06
I x C	8.21	<0.005	0.30	22.03	<0.001	0.54	5.46	<0.05	0.22	1.95	0.18	0.09

Note. Degrees of freedom: Omnibus. Item: 1,19; Condition, I x C: 2, 38
Pairwise. Item, Condition, I x C: 1,19

Table 2. Peak MNI coordinates with Brodmann's Area (BA) and peak F-statistic for the omnibus ANOVA including item (probe, irrelevants) and condition (NK, CK and CM) as repeated measure factors.

Effect	Regions	BA	MNI Coordinates			F
			x	y	z	
Item (I)						
	R Supramarginal Gyrus	40	60	-39	28	40.96
	R Supramarginal Gyrus	40	52	-42	25	38.27
	R Angular Gyrus	39	56	-49	30	30.36
	R Medial Frontal Gyrus	8	45	6	36	34.44
	R Precuneus	7	8	-64	36	30.11
	R Precuneus	7	15	-67	44	20.04
	R Precuneus	7	12	-52	49	18.96
	R Supp Motor Area	6	12	23	57	29.97
	R Supp Motor Area	8	3	20	55	25.56
	R Superior Frontal Gyrus	6	18	9	64	25.15
	R Middle Cingulate Cortex	23	2	-27	27	29.39
	R Middle Cingulate Cortex	23	6	-33	25	17.16
	L Supramarginal Gyrus	40	-54	-52	30	28.20
	L Supramarginal Gyrus	40	-58	-31	37	26.90
	R Inferior Frontal Gyrus Pars Opercularis	44	51	15	3	22.55
	R Inferior Frontal Gyrus Pars Orbitalis	47	50	26	-9	20.87
	R Insula	45	45	21	1	19.79
Condition (C)						
	L Supp Motor Area	8	2	20	52	20.38
	L Supp Motor Area	8	3	23	45	18.12
	L Superior Medial Frontal Gyrus	9	2	42	25	13.34
	R Superior Temporal Gyrus	39	57	-49	21	11.04
	R Middle Temporal Gyrus	39	50	-52	20	10.65
	R Middle Temporal Gyrus	39	44	-55	15	8.81
I x C						
	R Inferior Frontal Gyrus Pars Orbitalis	47	51	21	-5	17.12
	R Inferior Frontal Gyrus Pars Triangularis	45	51	24	10	12.77
	R Insula	13	38	23	3	12.32
	L Inferior Frontal Gyrus Pars Orbitalis	47	-44	20	-5	16.29
	L Insula	13	-36	24	1	12.21
	R Middle Frontal Gyrus	10	34	48	12	15.41
	R Superior Medial Gyrus	8	2	24	45	13.18
	R Anterior Cingulate Cortex	32	6	38	16	11.84
	R Anterior Cingulate Cortex	32	9	47	9	11.14
	R Middle Temporal Gyrus	37	56	-45	4	12.11
	R Middle Temporal Gyrus	39	45	-54	14	11.04
	R Superior Temporal Gyrus	39	57	-46	18	9.76

Note: Significance at all regions for each contrast was tested with an F test at $p < 0.001$ at the voxel level (uncorrected), with $p < 0.05$ (FWE-corrected) at the cluster level.
Degrees of freedom: Item: 1, 114; Condition, Ix C: 2, 114. BA, Brodmann's area; L, Left Hemisphere; R, Right Hemisphere.

Table 3. Peak MNI coordinates with Brodmann's Area (BA) and minimal t-statistic of brain regions with significant positive and negative responses in the contrast of Probe minus Irrelevants in the NK, CK and CM conditions. For the NK condition, the probe was only nominally defined because participants had no concealed information.

Condition	Regions	BA	MNI Coordinates x	y	z	t
	Probe > Irrelevants					
NK						
	-	-	-	-	-	-
CK						
	R Middle Cingulate Cortex	24	4	21	37	6.98
	R Anterior Cingulate Cortex	32	6	40	16	5.46
	L Superior Medial Frontal Gyrus	8	-2	26	57	5.24
	L Inferior Frontal Gyrus Pars Orbitalis	47	-46	15	-3	6.63
	L Inferior Frontal Gyrus Pars Triangularis	47	-34	26	-2	4.74
	L Insula	47	-27	21	-3	4.31
	R Precuneus	7	10	-66	36	6.39
	R Inferior Parietal Lobule	40	50	-45	48	5.63
	R Angular Gyrus	39	44	-55	39	5.61
	R Supramarginal Gyrus	40	51	-39	24	5.56
	R Caudate	-	9	3	3	5.43
	R Pallidum	-	15	-3	-2	5.20
	R Inferior Frontal Gyrus Pars Triangularis	45	52	22	9	5.06
	R Inferior Frontal Gyrus Pars Orbitalis	47	48	27	-8	4.86
CM						
	R Supramarginal Gyrus	40	63	-42	27	5.40
	R Supramarginal Gyrus	40	52	-42	25	3.90
	Irrelevants > Probe					
NK						
	-	-	-	-	-	-
CK						
	-	-	-	-	-	-
CM						
	R Gyrus Rectus/Medial Orbitofrontal Cortex	11	2	42	-18	4.75

Note: Significance at all regions for each contrast was tested by a one-sample t-test at $p < 0.001$ at the voxel level (uncorrected), with $p < 0.05$ (FWE-corrected) at the cluster level.
BA, Brodmann's area; L, Left Hemisphere; R, Right Hemisphere.

Table 4. Peak MNI coordinates and minimal t-statistic of brain regions showing significant probe effect differences between conditions (CK-NK, CM-NK, and CK-CM).

Condition	Regions	BA	MNI Coordinates			t
			x	y	z	
CK - NK						
	L Insula	48	-36	17	4	7.04
	L Inferior Frontal Gyrus Pars Orbitalis	48	-45	15	-5	6.63
	L Inferior Frontal Gyrus Pars Oercularis	48	-50	11	1	3.97
	L Superior Medial Frontal Gyrus	8	2	24	45	6.49
	R Inferior Frontal Gyrus Pars Triangularis	45	48	21	9	5.63
	R Superior Frontal Gyrus	10	33	51	10	5.55
	R Pallidum	-	18	-1	0	5.49
	R Caudate	-	15	12	6	4.31
	R Inferior Parietal Lobule	40	48	-45	48	5.44
CM - NK						
	-	-	-	-	-	-
CK - CM						
	-	-	-	-	-	-

Note: Significance was tested with a paired t-test at $p < 0.001$ at the voxel level (uncorrected), with $p < 0.05$ (FWE-corrected) at the cluster level.
BA, Brodmann's area; L, Left Hemisphere; R, Right Hemisphere.

Table 5. ANOVAs comparing activation for the probe and the irrelevants across 7 ROIs and conditions (omnibus, CK-NK-CM; condition pairs, CK-NK, CM-NK, CK-CM). For the NK condition, the probe was only nominally defined because participants had no concealed information.

	Omnibus			Pairwise								
	CK - NK - CM			CK - NK			CM - NK			CK - CM		
Source	F	p	η_p^2	F	p	η_p^2	F	p	η_p^2	F	p	η_p^2
Item	19.84	<0.001	0.51	14.28	<0.001	0.43	4.89	<0.05	0.21	26.26	<0.001	0.58
Condition	9.62	<0.005	0.34	14.88	<0.001	0.44	7.71	<0.05	0.29	2.99	0.1	0.14
ROI	1.26	0.29	0.06	0.88	0.49	0.04	0.82	0.52	0.04	2.19	0.07	0.10
I x C	11.91	<0.001	0.39	21.99	<0.001	0.54	6.36	<0.05	0.25	6.08	<0.05	0.24
I x R	1.95	0.11	0.09	0.78	0.54	0.04	1.47	0.23	0.07	2.70	<0.05	0.12
C x R	1.68	0.13	0.08	1.95	0.12	0.09	2.30	0.06	0.11	1.00	0.41	0.05
I x C x R	1.24	0.29	0.06	1.38	0.25	0.07	1.89	0.13	0.09	0.41	0.78	0.02

Note. Degrees of freedom: Omnibus. Item, Condition, I x C: 1, 19; ROI, I x R, C x R, I x C x R: 6, 114. Pairwise. Item: 1, 19; Condition, I x C: 2, 38; ROI, I x R: 6, 114; C x R, I x C x R: 12, 228.

Table 6. Area under the curve (AUC) for the comparisons CK vs NK and CM vs NK for 7 ROIs.

	ROI							
Comparison	LGFi	LLPi	GC	RGFi	RLPi	Thal	GFd	Mean
CK vs NK	**0.85***	**0.76**	**0.74***	**0.82**	**0.81**	**0.73***	**0.83****	**0.86***
CM vs NK	**0.69**	0.66	0.57	**0.74**	**0.71**	0.56	0.61	0.71

Note: LGFi: left inferior frontal gyrus/insula, LLPi: left inferior parietal lobule/supramarginal gyrus, GC: middle cingulate gyrus, RGFi: right inferior frontal gyrus/insula, RLPi: right inferior parietal lobule/supramarginal gyrus, Thal: thalamus/caudate nucleus/lenticular nucleus, GFd: medial frontal gyrus/superior frontal gyrus.

AUC numbers in bold indicate values significantly above chance (0.5). The asterisks indicate a significant difference between the CK/NK and CM/NK AUCs: * $p<.05$, ** $p<.005$. The last column shows the results for mean ROI activation.

Table 7. Area under the curve (AUC) for the independent CK-NK and CM- NK analyses and for the CK-CM generalization analyses for the ROI and whole-brain datasets

	Analysis		
Comparison	Multiaverage ROI	Multivoxel ROI	Multivoxel whole brain
CK/NK	**0.85****	**0.83****	**0.80**
CM/NK	0.63	0.63	**0.79**
Generalization			
CK/NK → CK/NK	**0.84***	**0.84***	**0.79**
CK/NK → CM/NK	**0.71**	**0.68**	**0.73**

Note: AUC numbers in bold indicate values significantly above chance (0.5). The asterisks indicate a significant difference between the CK/NK and CM/NK AUCs: * $p<.05$, ** $p<.005$.